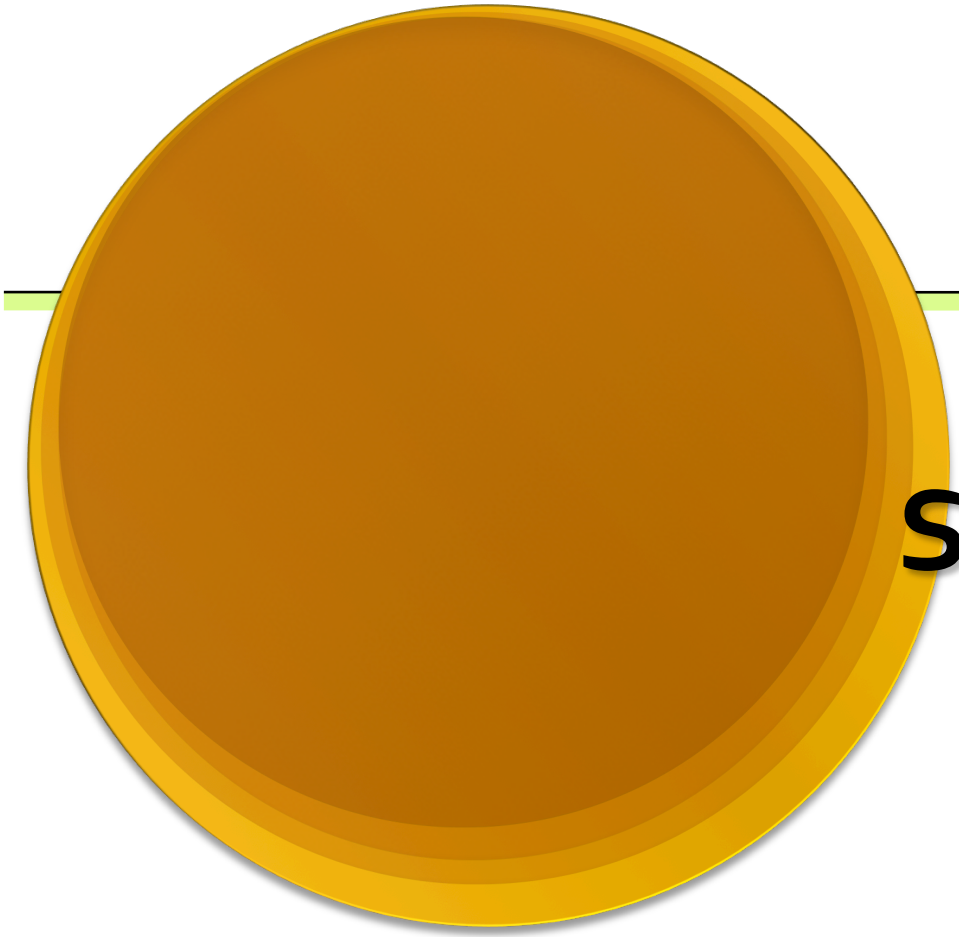


ST2334 2011/2012 Semester 2

Topic 1

Introduction to Statistics

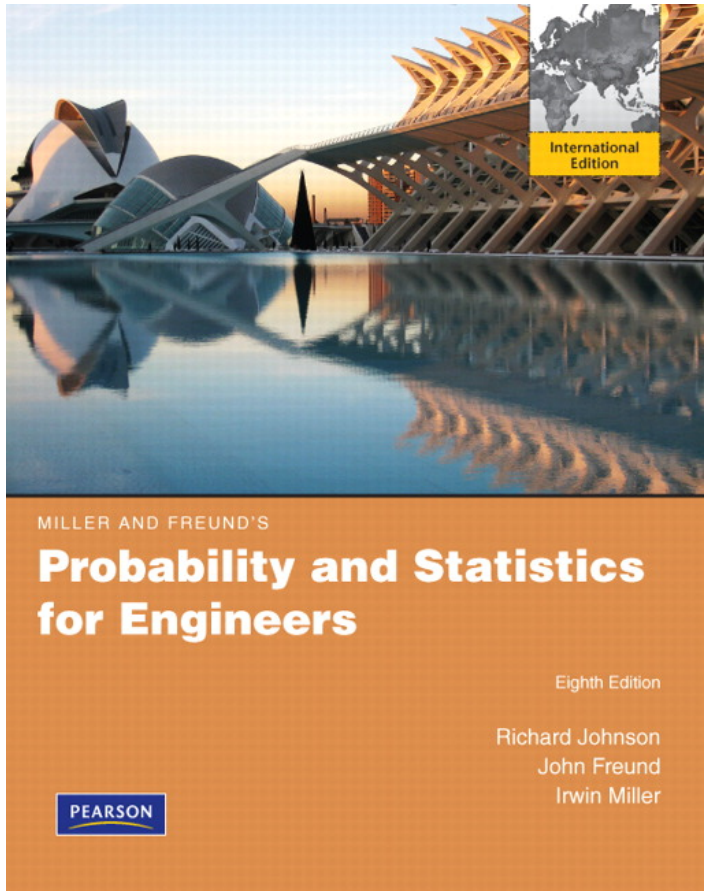
ST2334 PROBABILITY AND STATISTICS

Miller & Freund's Probability and Statistics for Engineers: International Edition, 8/e

Author : Richard A. Johnson

Publisher : Pearson

ISBN : 9780321694980



Available at NUS Co-op @ LT 27 !!

What is Statistics?

- **Merriam-Webster Dictionary**
  - **sta·tis·tics, *noun pl but singular or pl in constr*** \stə-ˈtis-tiks\
  - **Def: a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data**
- **Two main parts**
  - **Descriptive Statistics**
    - **Summarizing and describing collected data**
  - **Statistical Inference**
    - **Drawing conclusions or making decisions about a population based on a sample**

Statistics in 4 Steps

- Set clearly defined goals for the investigation.
- Make a plan of what data to collect and how to collect it.
- Apply appropriate statistical methods to extract information from the data.
- Interpret the information and draw conclusions.

Basic terms in Statistics

Variable	A characteristic of research interest
Unit	An item or subject being measured
Observation	Information recorded from one unit
Population units	Complete collection of units
Statistical Population	Complete collection of observations
Sampling units	Partial collection of units
Sample	Partial collection of observations

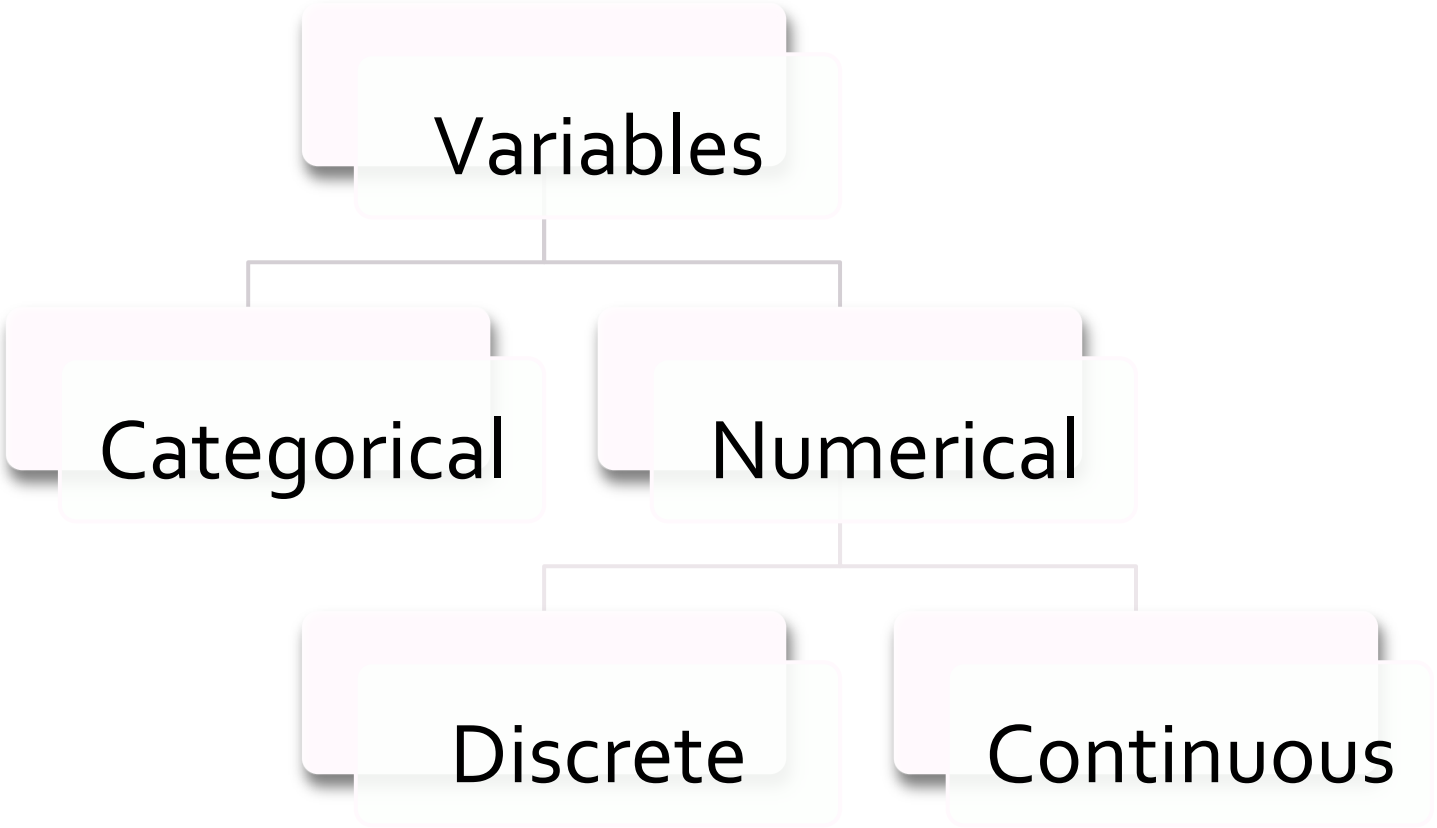
Example: BMI of NUS students

- Variable — BMI
- Unit — NUS student
- Observation — NUS student's BMI
- Population units — All NUS students
- Statistical Population — All BMIs of NUS students
- Sampling units — Some NUS students
- Sample — Some BMIs of NUS students

Example: BMI of NUS students

- Descriptive statistics
  - Collect all BMI data. Compute average, how many percent obese etc. More tools later.
- Statistical inference
  - Can estimate average, how many percent obese etc. based on sample.
  - Relationship with other variables? Blood pressure, CAP?

Types of Variables



Variable Examples

- **Categorical** — Race, Gender, Hair color, Field of study, favourite Big Bang Theory character
- **Discrete** — Age, Pairs of shoes owned, number of friends on Facebook
- **Continuous** — Weight, Height, Temperature, Distance from hall to LT27

Polio Experiment

- In early 1900s, Polio claimed hundreds of thousands of victims especially children. By 1950s, several vaccines had been discovered. How do we know if they're effective?
- Want to study the relationship between vaccination and incidence. Experiment!
- Give vaccine to everyone, incidence of polio went down.
  - Does that show the vaccine is effective?

Polio: Controls

- No! incidence rate vary from year to year.
  - E.g. 60k cases in '52, 30k cases in '53.
- We need to compare our results with that of an untreated group – *control group* (or just controls).
  - This is example of using *historical* controls.
- If possible we should use *contemporaneous* controls.
  - Idea: since children could only be vaccinated with parents' permission, we can use the no-consent group as controls!

Polio: Confounder

- Bad idea!
  - Higher income parents are more likely to consent
  - AND children from higher income family more likely to get polio (true story)
- When 2 groups differ with respect to some factor other than treatment AND the factor affects the treatment, we say the effect is *confounded* with the treatment.
- In this case, income is a *confounder*.

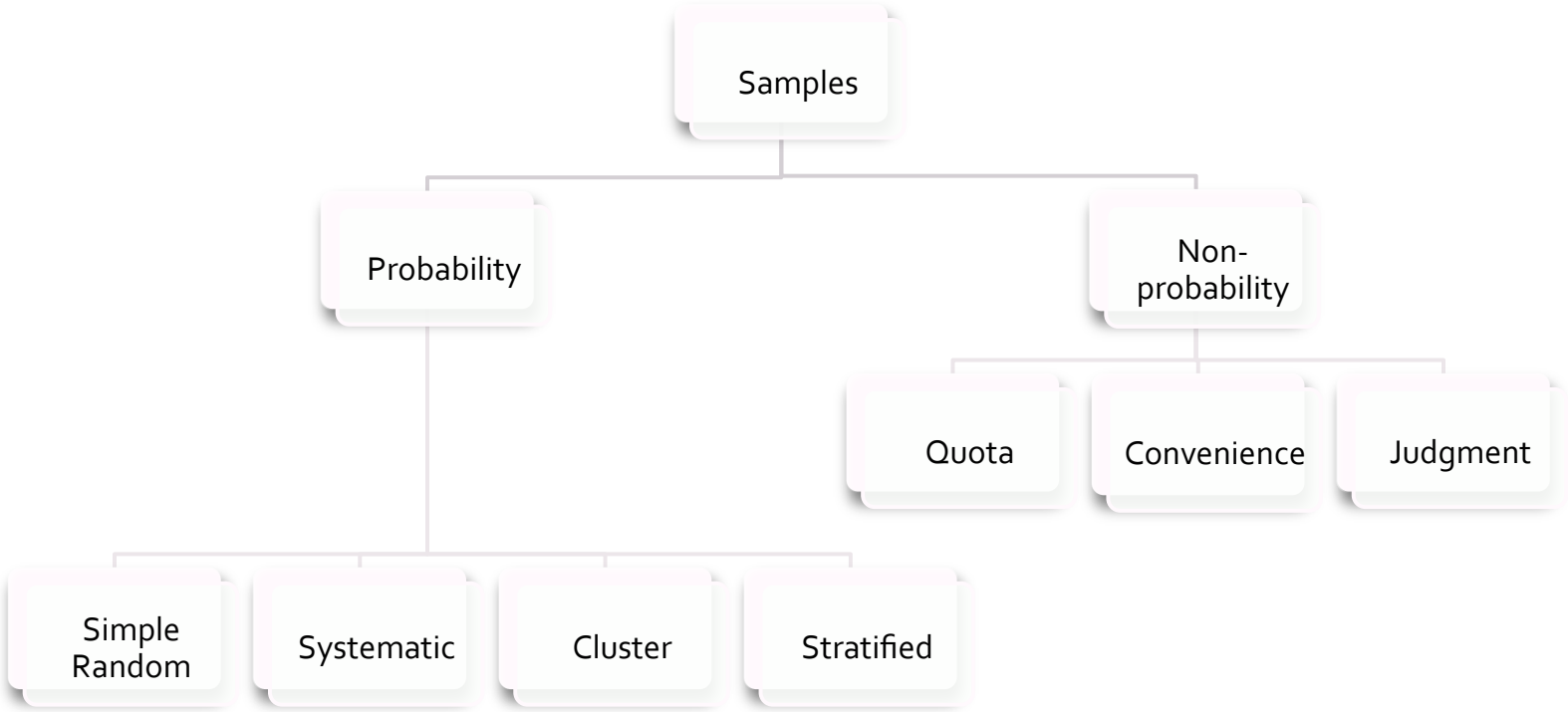
Gold standard experiment

- Randomized Controls
  - Treatment and control groups are randomly assigned
  - For the polio example, randomized those with consent
- Double blind
  - Placebo effect well documented.
    - Patients need to be blind to whether they receive treatment
  - Investigators should be blind too
    - Polio is hard to diagnose; diagnosis could be affected (perhaps subconsciously) by knowing if patient is vaccinated.

Smoking Study

- Sometimes not possible to do experiments.
  - Assign subject to smoke to investigate if smoking causes cancer?!
- We can instead do an *observational study*.
  - We observe and see that smokers are more likely to get lung cancer than non-smokers.
  - Conclude that there is strong association between smoking and lung cancer.
- Association is not causation.

Types of Samples



Convenience Sampling

- What proportion of NUS students play Angry Birds?
- Ok who plays Angry Birds raise your hands.
- Convenient indeed!
- Representative of population?

Quota Sampling

- Split population into mutually exclusive groups, and sample them proportionally.
- Perhaps CS majors more likely to play Angry Birds, so let's split by major. For e.g., if 5% of NUS population major in CS, we require 5% of the sample to major in CS.
- No other requirement. Pick all guys?
- Quota that too?
- Still have to be careful.

Judgment Sampling

- Make your own judgment to select what is an appropriate sample.
- Let's stand outside Yusof Ishak Hall during lunch break. Looks random enough.
- Good enough?
- Even expert judgment may unwittingly introduce bias.

Simple Random Sample

- Draw each sample unit independently at random.
  - Each unit has equal chance of being sampled
  - Any subset of k individuals has the same chance of being sampled as any other subset of k individuals. (for all k)
- E.g. Assign every NUS student a number 1,2,3,… Use a random number table.
- The gold standard, but often costly and/or difficult to implement.

Systematic Sampling

- Ordered list, random starting point, take every kth member afterwards.
- E.g. Sort matriculation card number, draw a random number from 1 to 10, take every 10th person.
- Need to ensure that the ordered list does not hide pattern that may affect outcome.

Stratified Sampling

- Divide units into mutually exclusive subgroups.
- Subgroups should have common properties. E.g. by major
- Use simple random sampling to choose observations from each subgroup to form sample.
- This is like quota sampling but makes use of chance instead of judgment.

Cluster Sampling

- Divide population into several subgroups
- Each subgroup is representative of the population
- Randomly select clusters
- Include all units from selected clusters to form sample
- By Hall?
- Better example: Dengue prevention inspection, inspect every flat in a few randomly selected hdb blocks

Sampling recap

- Non-probability samples are often more convenient to obtain.
  - But need to be mindful of pitfalls.
- Probability samples involve planned use of chance.
  - Eliminates subjectivity
  - More "likely" to give us a representative sample. No guarantees, we will learn to quantify how likely.